REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
June 2016	Final/Technical	

4. TITLE AND SUBTITLE
Rich Representations with Exposed Semantics for Deep Visual Reasoning

5a. CONTRACT NUMBER
N00014-10-1-0934

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Davis, Larry; Chellappa, Rama; Hoiem, Derek; Gupta, Abhinav; Hebert, Martial; Aminoff, Elissa; Park, HyunSoo; Forsyth, David; Shi, Jianbo; Tarr, Michael

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Prepared in cooperation with subawards: University of California Berkeley, University of Maryland, University of Pennsylvania, University of Illinois at Urbana-Champaign, University of Washington

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Daniel Joyner
					19b. TELEPHONE NUMBER *(Include area code)*
U		U		10	412-268-5730

Contract Number: N000141010934

Title: Rich Representations with Exposed Semantics for Deep Visual Reasoning

Major Goals:

The objective of this MURI is to develop techniques that can explain complex images and videos in common sense terms. The emphasis is on how to acquire flexible visual representations that can be shared across tasks and interpreted by humans. Our approach to representations addresses the challenges of describing unfamiliar objects, scenes, activities by exploiting shared properties, and by including the complex interactions. Our reasoning tools emphasize efficient approaches for dealing with the complex structure of the world, focused reasoning to reason about the relevant parts of the visual input, and temporal reasoning to deal with events. We aim to develop approaches to visual reasoning that can incorporate the constraints of the specific task at hand and the need to present useful and relevant information to a human. In addition, we explore in detail the design of the datasets used for training visual reasoning elements and the limitations of these datasets; and we explore connections between the computer vision aspects of the work and human vision studies in cognitive neuroscience.

Accomplishments Under Goals:

Representing visual knowledge.

We continued our research on binary coding of images and videos started in the previous years of the Grant. We presented an efficient algorithm for supervised hashing which updates hash functions as classes are added go or subtracted from the set of classes represented in a database. We showed our hash functions could be learned that better preserve the semantics of the class.
Work continued on a subset selection approach for choosing the most useful attributes for real-world visual recognition problems. Low-level features do not adequately characterize the semantic content in images, or the spatio-temporal structure in videos. In this work, we consider two types of attributes. One type of attributes is generated by humans, while the second type is data-driven attributes extracted from data using dictionary learning methods. Experimental results on four public datasets demonstrate that the proposed attribute-based representation significantly boosts the performance of visual recognition algorithms.

Reasoning with visual knowledge.

In the area of optimization and learning tools for reasoning, we continued work on exploiting submodular optimization theory for computer vision problems. Specifically, we considered the problem of selecting best attributes for human action recognition. Experimental results on the Olympic Sports and UCF101 datasets demonstrate that the proposed attribute-based representation can significantly boost the performance of action recognition algorithms and outperform most recently proposed recognition approaches. Building on the CNN work started last year, we have also proposed an approach for unsupervised learning of ConvNets from Images and Videos. Using this approach, we have come within 2.5% of ImageNet performance on VOC 2012. For images, we use supervision from relative location of patches. With this supervision, we train a siamese double tower network. For video supervision, we track patches and use the patches from beginning and end of the track to train an end-to-end deep network. We have also proposed an approach to train ConvNets from Noisy labeled data using a curriculum-learning approach. This network performed surprisingly better than ImageNet network; eliminating the need for labelled large datasets.
We have made progress in the area of object detection and scene understanding. Specifically, we continued the work on regression trees and regression forest framework with applications in pose estimation and vehicle direction estimation. The proposed methods outperform all the baseline regression methods.
To complement the work on detection techniques, we investigated new approaches to reasoning about focus of attention and spatial localization. Specifically, we demonstrated a technique that learns which regions in an image are relevant for a given question and possible answer. We showed that, by learning where to look as a latent task, the system substantially outperforms methods that attempt to directly predict the answer from the entire image. We demonstrated a technique to learn latent landmarks that help localize hard-to-see parts by using scene context. For example, the system learns to find nearly invisible car door handles by first finding the wheel, then the corner of the window, then finally the handle.
In the area of scene understanding, in collaboration with MERL researchers, we developed an approach for semantic segmentation using Gaussian networks. By combining the proposed GMF network with deep Convolutional Neural Networks (CNNs), we propose a new end-to-end trainable Gaussian conditional random

field network. The proposed GCRF network outperforms various recent semantic segmentation approaches that combine CNNs with discrete CRF models.
A further focus has been to learn methods to estimate what the objects are made of. We investigated methods for learning a cost function such that the solution to the cost function is an intrinsic image (e.g., albedo, material or shading properties). These methods are very demanding of training data, and we demonstrated that they work well in automated colorization of monochrome images (where data is easily available).
Finally, we continued our work on reasoning techniques that are specific to actions and dynamic environments. In particular, research continued on developing representations and algorithms for recognition of human actions using 3D skeletal data. We explored the notion of rolling maps, a well-defined mathematical concept that has not been explored much by the computer vision community. Experimental results on two challenging action datasets show that the proposed approach performs equally well or better when compared to the state-of-the-art. We also expanded our research on perceiving the interactions between people. We use first-person video to obtain in-situ measurements of these natural interactions. We implemented two instances of this concept two examples in the papers published. (1) At physical level, we predict the wearer's intent in a form of force and torque that control the movements. (2) At spatial scene level, we predict plausible future trajectories of ego-motion. The predicted paths avoid obstacles, move between objects, even turn around a corner into invisible space behind objects.

Acquiring and manipulating visual knowledge.

A major advance this year is the demonstration of how large datasets can be acquired in the context of robotics tasks, specifically robot grasping tasks. This work paves the way toward coupled learning of visual and action in an unsupervised manner. In this part of the work, we take the leap of increasing the available training data leading to a dataset size of 50K data points collected over 700 hours of robot grasping attempts. We showed how this allowed us to train a Convolutional Neural Network (CNN) for the task of predicting grasp locations. Our experiments clearly show the benefit of using large-scale datasets (and multi-stage training) for the task of grasping.
Indetection and recognition, we showed how discriminative triplets of patches could be mined from big visual data, capturing geometric constraints on appearance and resulting in better fine grained recognition. We showed how a search strategy based on reinforcement learning could be learned from a large database of labeled images to control the efficient search for specific object classes in images.
Finally, we showed how large, unlabeled data can be used effectively in learning for correspondence-related tasks. This part of the work tackles the problem of establishing dense visual correspondence across different object instances. For this task, although we do not know what the ground-truth is, we know it should be consistent across instances. We exploit this consistency as a supervisory signal to train a convolutional neural network to predict cross-instance correspondences. We demonstrated that our end-to-end trained ConvNet supervised by cycle-consistency outperforms state-of-the-art pairwise matching methods in correspondence-related tasks.

Neuroscience.

we explored how the neural mechanism underlying scene understanding includes the processing of contextual associations that links past experiences with the current perceptual input. This year, we addressed this by recording and stimulating neurons in a single human patient who had implanted electrodes due to suffering from intractable epilepsy. This provides critical evidence of a relationship between visual recognition, associative processing, and episodic memory and provides important clues into the neural mechanism underlying scene understanding. In addition to this study, we are following up on a previous finding produced under this MURI in which, using fMRI we found our computer vision model - NEIL, which is learns mid-level visual features from millions of real-world images - can account for aspects of processing in scene understanding.

Training Opportunities:

Results Dissemination

Plans Next Reporting Period

Nothing to Report

Honors and Awards

Abhinav Gupta - Sloan Research Fellow
Abhinav Gupta and Lerrel Pinto - ICRA 2016 Best Student Paper award
Abhinav Gupta - IJCAI Early Career Spotlight

Protocol Activity Status

Distribution Statement:

Approved for public release; distribution is unlimited.

Participants

First Name: Larry **Last Name:** Davis

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Rama **Last Name:** Chellappa

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Derek **Last Name:** Hoiem

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 2

Countries of Collaboration

First Name: Abhinav **Last Name:** Gupta

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Martial **Last Name:** Hebert

Project Role: PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Elissa **Last Name:** Aminoff

Project Role: Staff Scientist (doctoral level)

National Academy Member: N **Months Worked:** 3

Countries of Collaboration

First Name: HyunSoo **Last Name:** Park

Project Role: Postdoctoral (scholar, fellow or other postdoctoral position)

National Academy Member: N **Months Worked:** 12

Countries of Collaboration

First Name: David **Last Name:** Forsyth

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Jianbo **Last Name:** Shi

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

First Name: Michael **Last Name:** Tarr

Project Role: Co PD/PI

National Academy Member: N **Months Worked:** 1

Countries of Collaboration

Rich Representations with Exposed Semantics for Deep Visual Reasoning

ONR MURI Topic 6

Grant N000141010934

Annual Progress Report

June 2016

Carnegie Mellon University

University of California Berkeley

University of Maryland

University of Pennsylvania

University of Illinois at Urbana-Champaign

University of Washington

Objectives

This document summarizes progress in MURI program N000141010934 "Rich Representations with Exposed Semantics for Deep Visual Reasoning" led by Carnegie Mellon University, with University of Maryland, University of Pennsylvania, and University of Illinois at Urbana-Champaign. The present report covers the second year of the Grant up to and including August 2015.

The objective of this MURI is to develop techniques that can explain complex images and videos in common sense terms. The emphasis is on how to acquire flexible visual representations that can be shared across tasks and interpreted by humans. Our approach to representations addresses the challenges of describing unfamiliar objects, scenes, activities by exploiting shared properties, and by including the complex interactions. Our reasoning tools emphasize efficient approaches for dealing with the complex structure of the world, focused reasoning to reason about the relevant parts of the visual input, and temporal reasoning to deal with events. We aim to develop approaches to visual reasoning that can incorporate the constraints of the specific task at hand and the need to present useful and relevant information to a human. In addition, we explore in detail the design of the datasets used for training visual reasoning elements and the limitations of these datasets; and we explore connections between the computer vision aspects of the work and human vision studies in cognitive neuroscience.

Technical Accomplishments

Representing visual knowledge.

We continued our research on binary coding of images and videos started in the previous years of the Grant. We presented an efficient algorithm for supervised hashing which updates hash functions as classes are added go or subtracted from the set of classes represented in a database. We showed our hash functions could be learned that better preserve the semantics of the class.

Work continued on a subset selection approach for choosing the most useful attributes for real-world visual recognition problems. Low-level features do not adequately characterize the semantic content in images, or the spatio-temporal structure in videos. In this work, we consider two types of attributes. One type of attributes is generated by humans, while the second type is data-driven attributes extracted from data using dictionary learning methods. Experimental results on four public datasets demonstrate that the proposed attribute-based representation significantly boosts the performance of visual recognition algorithms.

Reasoning with visual knowledge.

In the area of optimization and learning tools for reasoning, we continued work on exploiting submodular optimization theory for computer vision problems. Specifically, we considered the problem of selecting best attributes for human action recognition. Experimental results on the Olympic Sports and UCF101 datasets demonstrate that the proposed attribute-based representation can significantly boost the performance of action recognition algorithms and outperform most recently proposed recognition approaches. Building on the CNN work started last year, we have also proposed an approach for unsupervised learning of ConvNets from Images and Videos. Using this approach, we have come within 2.5% of ImageNet performance on VOC 2012. For images, we use supervision from relative location of patches. With this supervision, we train a siamese double tower network. For video supervision, we track patches and use the patches from beginning and end of the track to train an end-to-end deep network. We have also proposed an approach to train ConvNets from Noisy labeled data using a curriculum-learning approach. This network performed surprisingly better than ImageNet network; eliminating the need for labelled large datasets.

We have made progress in the area of object detection and scene understanding. Specifically, we continued the work on regression trees and regression forest framework with applications in pose estimation and vehicle direction estimation. The proposed methods outperform all the baseline regression methods.

To complement the work on detection techniques, we investigated new approaches to reasoning about focus of attention and spatial localization. Specifically, we demonstrated a technique that learns which

regions in an image are relevant for a given question and possible answer. We showed that, by learning where to look as a latent task, the system substantially outperforms methods that attempt to directly predict the answer from the entire image. We demonstrated a technique to learn latent landmarks that help localize hard-to-see parts by using scene context. For example, the system learns to find nearly invisible car door handles by first finding the wheel, then the corner of the window, then finally the handle.

In the area of scene understanding, in collaboration with MERL researchers, we developed an approach for semantic segmentation using Gaussian networks. By combining the proposed GMF network with deep Convolutional Neural Networks (CNNs), we propose a new end-to-end trainable Gaussian conditional random field network. The proposed GCRF network outperforms various recent semantic segmentation approaches that combine CNNs with discrete CRF models.

A further focus has been to learn methods to estimate what the objects are made of. We investigated methods for learning a cost function such that the solution to the cost function is an intrinsic image (e.g., albedo, material or shading properties). These methods are very demanding of training data, and we demonstrated that they work well in automated colorization of monochrome images (where data is easily available).

Finally, we continued our work on reasoning techniques that are specific to actions and dynamic environments. In particular, research continued on developing representations and algorithms for recognition of human actions using 3D skeletal data. We explored the notion of rolling maps, a well-defined mathematical concept that has not been explored much by the computer vision community. Experimental results on two challenging action datasets show that the proposed approach performs equally well or better when compared to the state-of-the-art. We also expanded our research on perceiving the interactions between people. We use first-person video to obtain in-situ measurements of these natural interactions. We implemented two instances of this concept two examples in the papers published. (1) At physical level, we predict the wearer's intent in a form of force and torque that control the movements. (2) At spatial scene level, we predict plausible future trajectories of ego-motion. The predicted paths avoid obstacles, move between objects, even turn around a corner into invisible space behind objects.

Acquiring and manipulating visual knowledge.

A major advance this year is the demonstration of how large datasets can be acquired in the context of robotics tasks, specifically robot grasping tasks. This work paves the way toward coupled learning of visual and action in an unsupervised manner. In this part of the work, we take the leap of increasing the available training data leading to a dataset size of 50K data points collected over 700 hours of robot grasping attempts. We showed how this allowed us to train a Convolutional Neural Network (CNN) for the task of predicting grasp locations. Our experiments clearly show the benefit of using large-scale datasets (and multi-stage training) for the task of grasping.

In detection and recognition, we showed how discriminative triplets of patches could be mined from big visual data, capturing geometric constraints on appearance and resulting in better fine grained recognition.

We showed how a search strategy based on reinforcement learning could be learned from a large database of labeled images to control the efficient search for specific object classes in images.

Finally, we showed how large, unlabeled data can be used effectively in learning for correspondence-related tasks. This part of the work tackles the problem of establishing dense visual correspondence across different object instances. For this task, although we do not know what the ground-truth is, we know it should be consistent across instances. We exploit this consistency as a supervisory signal to train a convolutional neural network to predict cross-instance correspondences. We demonstrated that our end-to-end trained ConvNet supervised by cycle-consistency outperforms state-of-the-art pairwise matching methods in correspondence-related tasks.

Neuroscience.

We explored how the neural mechanism underlying scene understanding includes the processing of contextual associations that links past experiences with the current perceptual input. This year, we addressed this by recording and stimulating neurons in a single human patient who had implanted electrodes due to suffering from intractable epilepsy. This provides critical evidence of a relationship between visual recognition, associative processing, and episodic memory and provides important clues into the neural mechanism underlying scene understanding. In addition to this study, we are following up on a previous finding produced under this MURI in which, using fMRI we found our computer vision model – NEIL, which is learns mid-level visual features from millions of real-world images – can account for aspects of processing in scene understanding.